# Correlation and Simple Linear Regression

Shannon Pileggi

STAT 217

# OUTLINE

# Chocolate Consumption, Cognitive Function, and Nobel Laureates by Franz H. Messerli, M.D.

*New England Journal of Medicine* 367(**16**), Oct 18, 2012

- chocolate contains flavanols, which are loosely linked to improved cognitive function
- Is there a relationship between country's level of chocolate consumption and its population's cognitive function?
- explanatory variable: per capita yearly chocolate consumption (in kg)
- response variable: total number of Nobel laureates per 10 million person (through 2011)

# Chocolate Consumption, Cognitive Function, and Nobel Laureates by Franz H. Messerli, M.D.

*New England Journal of Medicine* 367(**16**), Oct 18, 2012

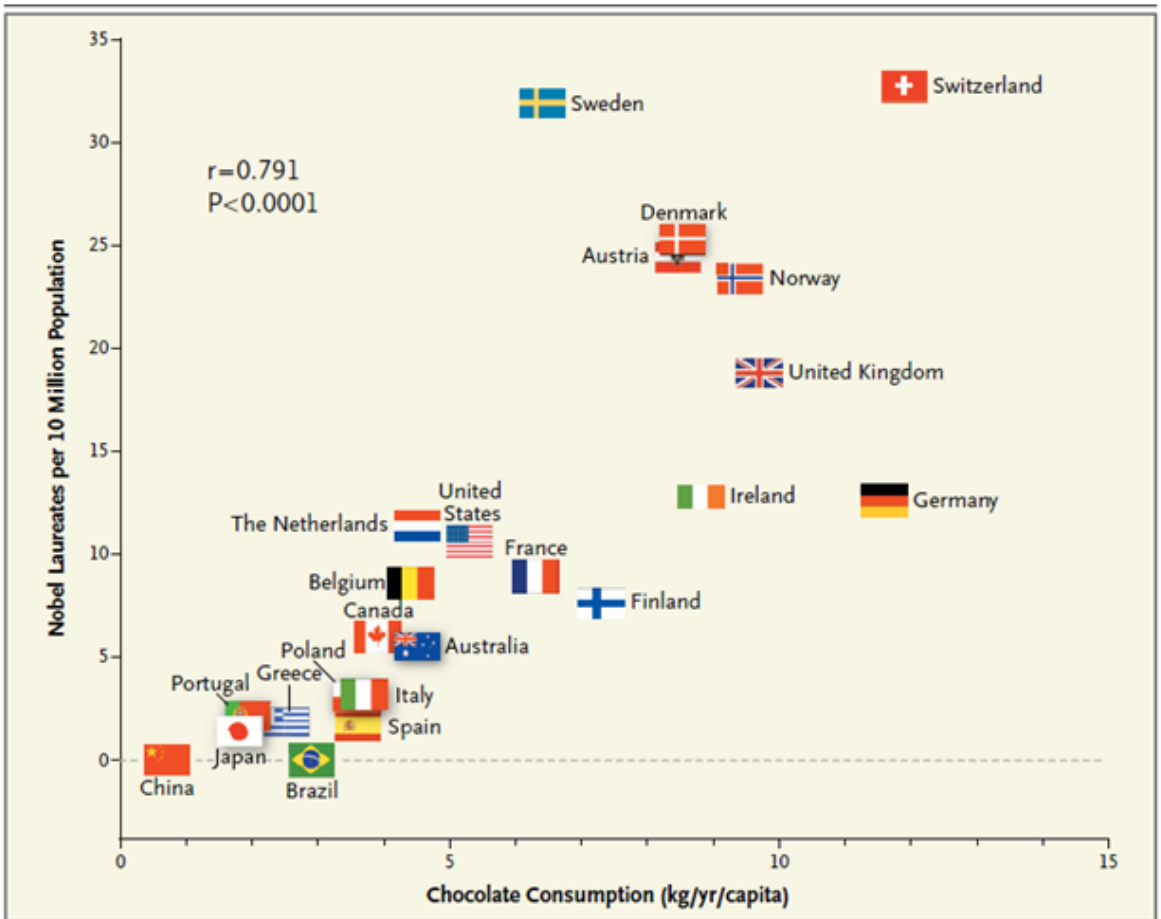



**Figure 1.** Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

# Discussion

- Is this study observational or experimental?
- What is the unit of observation?
- Describe the sample:
- Describe the population:
- What are other variables may affect the relationship between chocolate consumption and nobel laureates?

# Discussion, continued

Other oddities:

- Does it make sense to use nobel laureates as a measure of a country's cognitive function?
- Do we even know if the nobel laureates themselves ate chocolate?
- Were the variables measured on the same temporal scale?
- How are nobel prize winners identified with a specific country?

Stats.org: Cacao or Caca? How the media bit into chocolate Nobel prize link

# In the media

- Time: Secret to Winning a Nobel Prize? Eat more Chocolate
- MSN: Want to win a Nobel Prize? Eat more chocolate
- Forbes: Chocolate and Nobel Prizes Linked in Study
- USA Today: Study links eating chocolate to winning Nobels
- Reuters: Eat chocolate, win the Nobel Prize?
- NPR: The Secret to Genius? It Might Be More Chocolate

# From the author (Messerli)

Reuters:

> *"I started plotting this in a hotel room in Kathmandu, because I had nothing else to do, and I could not believe my eyes," he told Reuters Health. All the countries lined up neatly on a graph, with higher chocolate intake tied to more laureates. The link was so strong it made a joke out of a statistic that virtually all studies in medical journals hinge on - the so-called p-value.*

NPR:

> *"I have published about 800 papers in peer-reviewed journals," he says, "and every single one of them stands and falls with the p-value. And now here I find a p-value of 0.0001, and this is, to my way of thinking, a completely nonsensical relation. Unless you or anybody else can come up with an explanation. I've presented it to a few of my colleagues, and nobody has any thoughts."*

# The Data

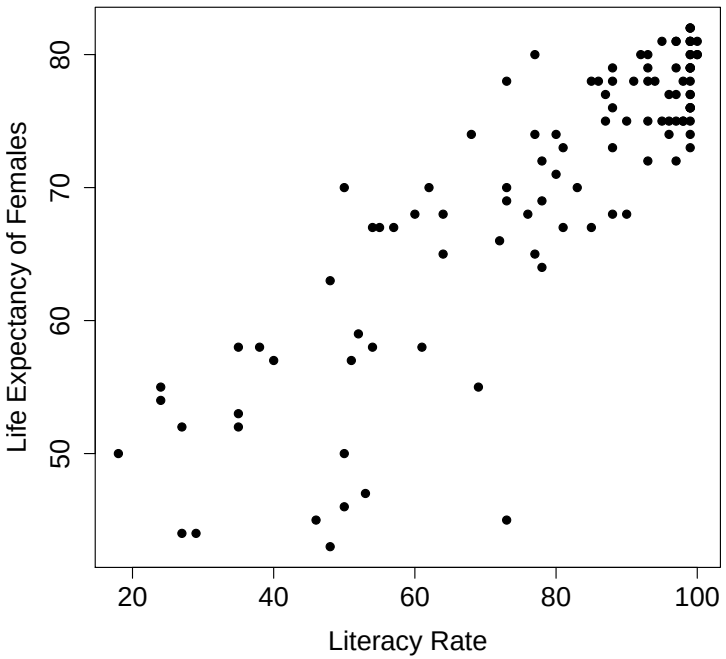
First 10 observations:

| | country | nobel_rate | chocolate | GDP_cap | totalpop | literate |
|---|---|---|---|---|---|---|
| 1 | Australia | 5.451 | 4.5 | 35052.512 | 22323900 | 99.9 |
| 2 | Austria | 24.332 | 10.2 | 36119.406 | 8423635 | 99.9 |
| 3 | Belgium | 8.622 | 4.4 | 33020.438 | 11047744 | 99.9 |
| 4 | Brazil | 0.050 | 2.9 | 10264.006 | 196935134 | 90.5 |
| 5 | Canada | 6.122 | 3.9 | 35738.703 | 34483975 | 99.8 |
| 6 | China | 0.060 | 0.7 | 7417.888 | 1344130000 | 95.4 |
| 7 | Denmark | 25.255 | 8.5 | 32601.660 | 5570572 | 99.9 |
| 8 | Finland | 7.600 | 7.3 | 32030.703 | 5388272 | 99.9 |
| 9 | France | 8.990 | 6.3 | 29963.223 | 65371613 | 100.0 |
| 10 | Germany | 12.668 | 11.6 | 34572.938 | 81797673 | 99.9 |

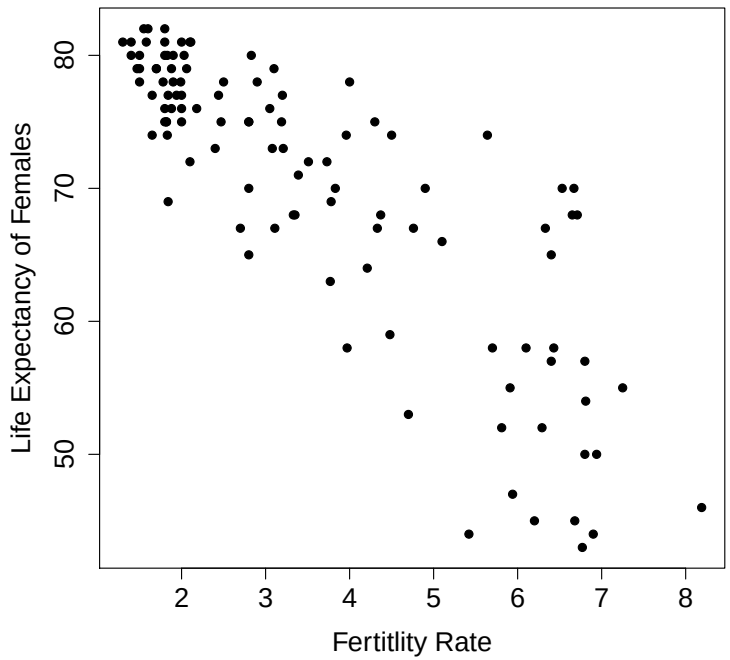
**Do `nobel_rate` and `chocolate` represent paired data?**

1. yes
2. no

# Association between two quantitative variables



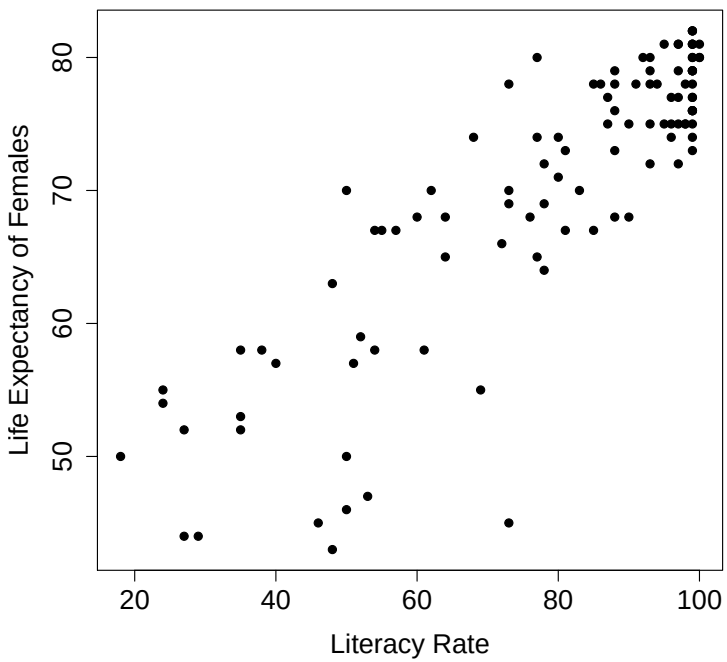

- Positive association: as literacy rate tends to increase, life expectancy of females also tends to increase



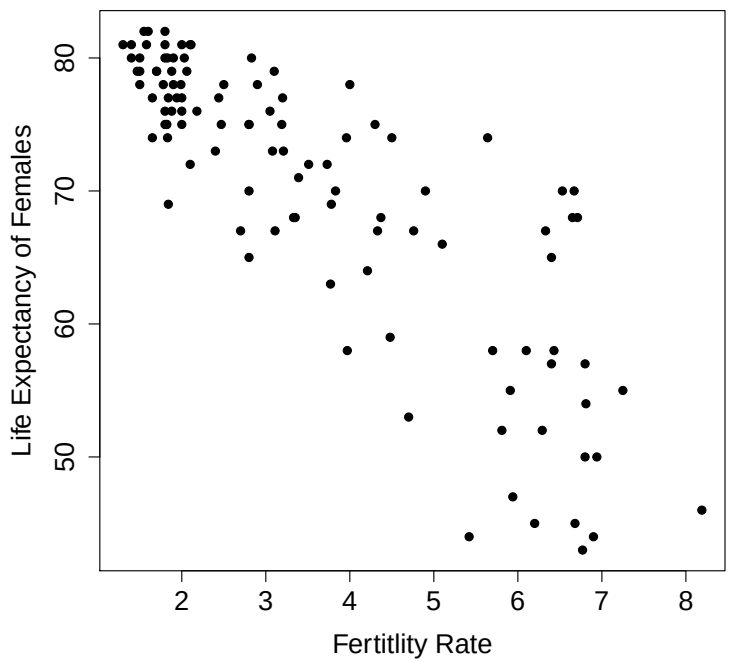

- Negative association: as fertility rate tends to increase, life expectancy tends to decrease

# Correlation

The **correlation** $r$ measures the **strength** and **direction** of the **linear association** between two quantitative variables. Correlation is always between between $-1$ and $+1$.

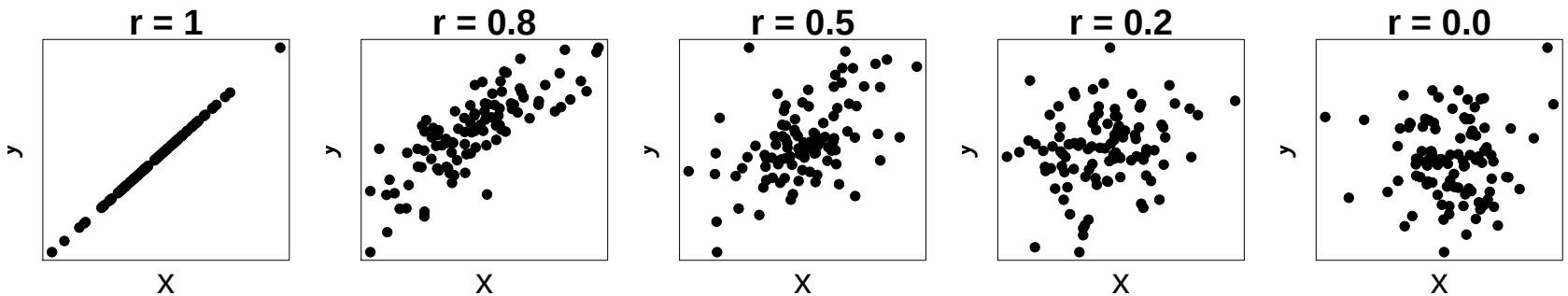



$r = 0.87$




$r = -0.84$

# Correlation examples




Rules of thumb:

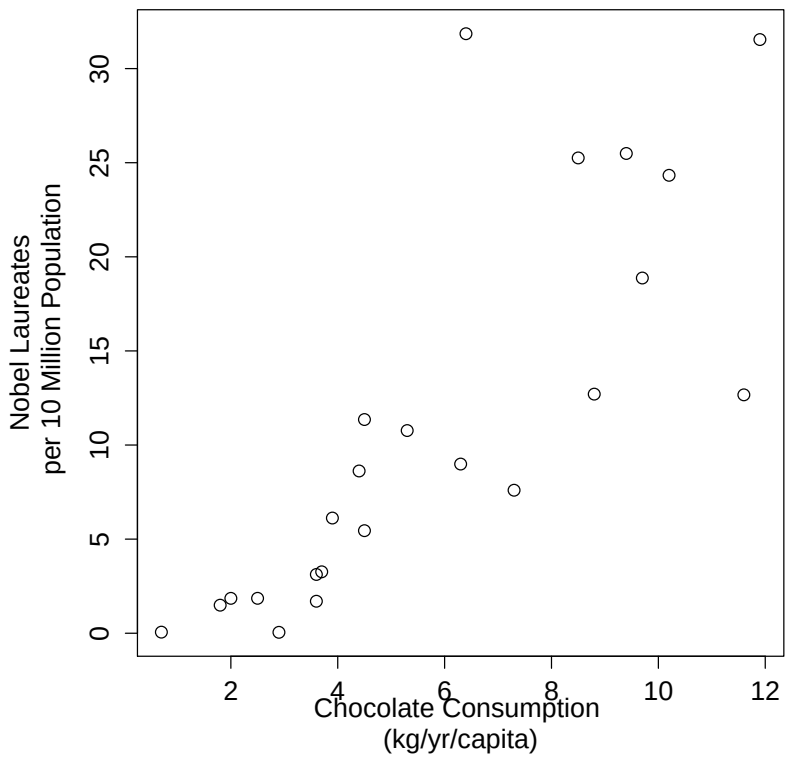
- $0 < |r| < 0.3$ - weak correlation
- $0.3 < |r| < 0.7$ - moderate correlation
- $0.7 < |r| < 1.0$ - strong correlation

**Which scenario illustrates the strongest linear association?**

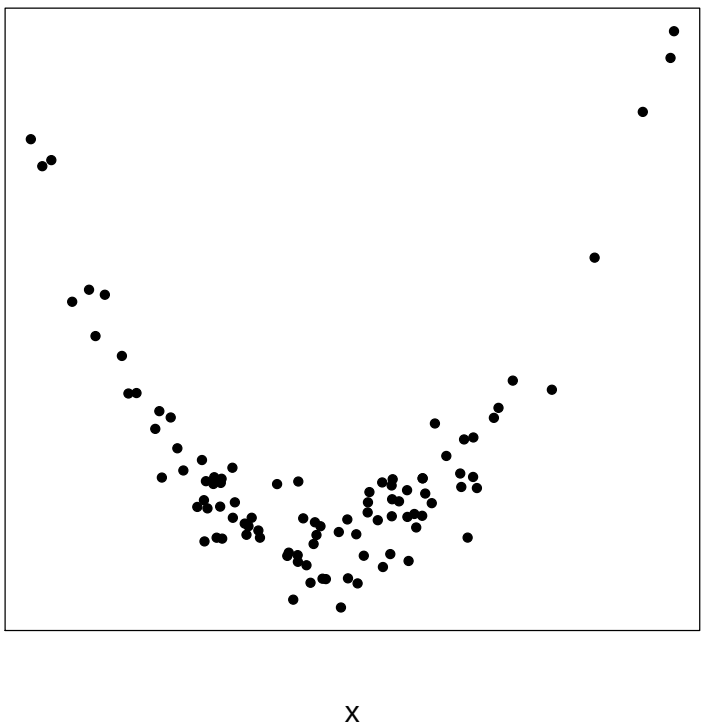
1. College GPA and high school GPA, $r = 0.46$
2. price and quality rating of bike helmets, $r = 0.30$
3. years of education and years in jail, $r = -0.53$
4. New York City marathon, age and finish time, $r = 0.04$
5. college GPA and a measure of tendency to procrastinate, $r = -0.36$

# Correlation estimate

```
> plot(NEJM$chocolate,NEJM$nobel_rate)
> cor(NEJM$chocolate,NEJM$nobel_rate)
[1] 0.8010949
```



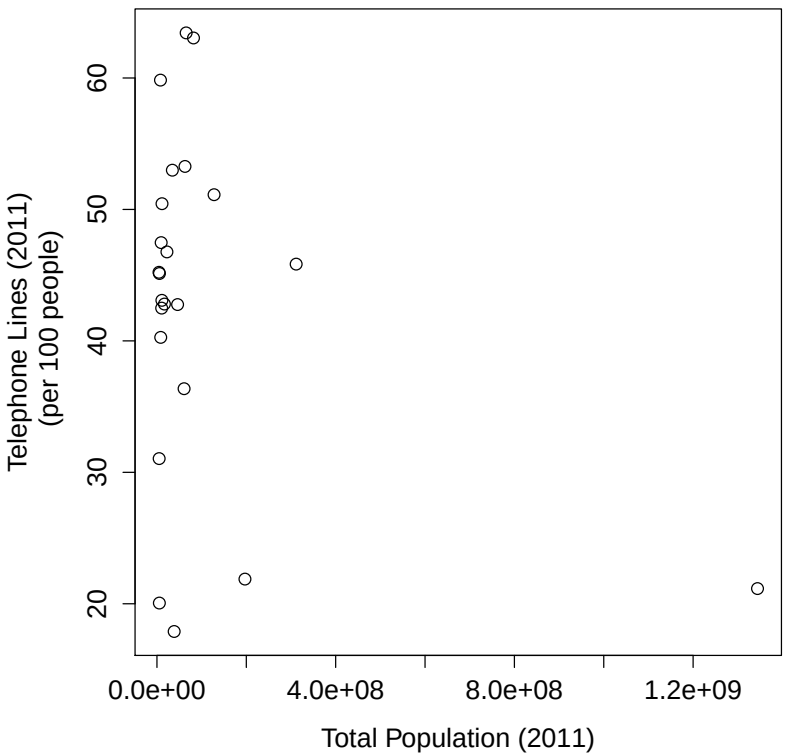
# Group Exercise




**What do you think the correlation between $x$ and $y$ is?**

1. $r = 1$
2. $r = 0.5$
3. $r = 0$
4. $r = -1$

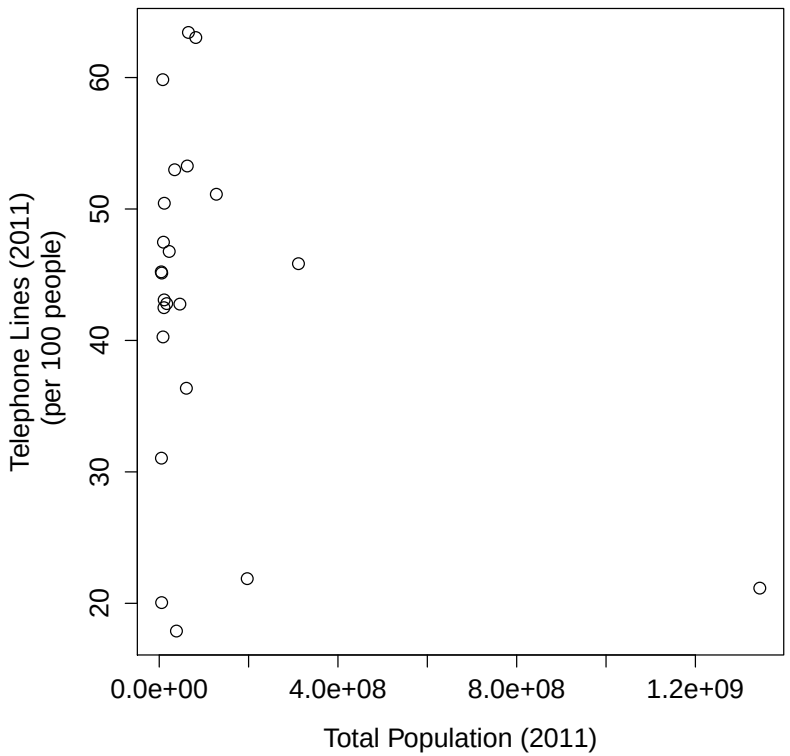
# Group Exercise



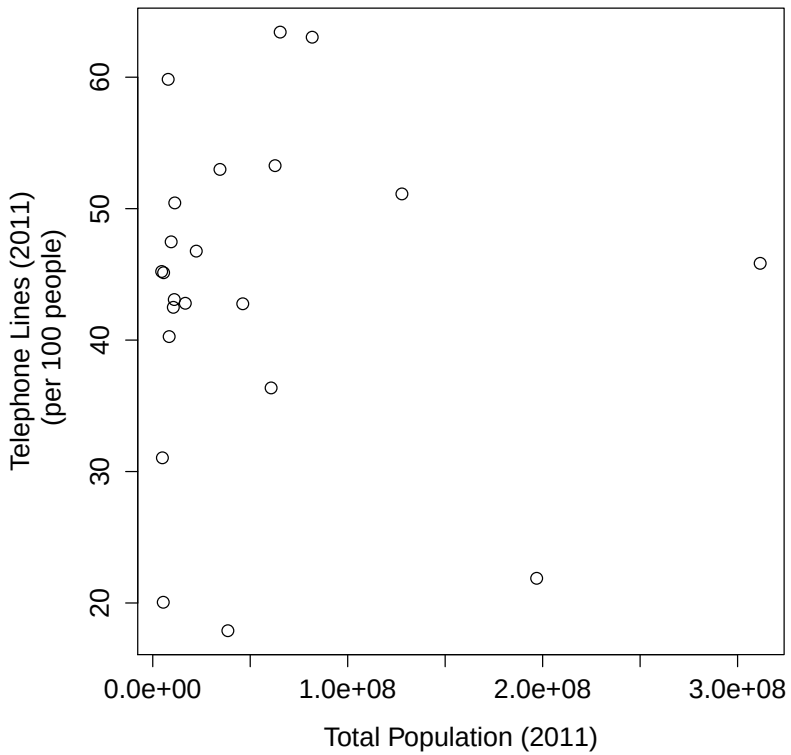

**Which of the following best describes the relationship between total population and telephone lines?**

1. there is a strong negative linear relationship
2. there is a moderate negative linear relationship
3. there is very little, if any, linear relationship
4. there is a moderate positive linear relationship
5. there is a strong positive linear relationship

# Chocolate Nobel Data Set Example

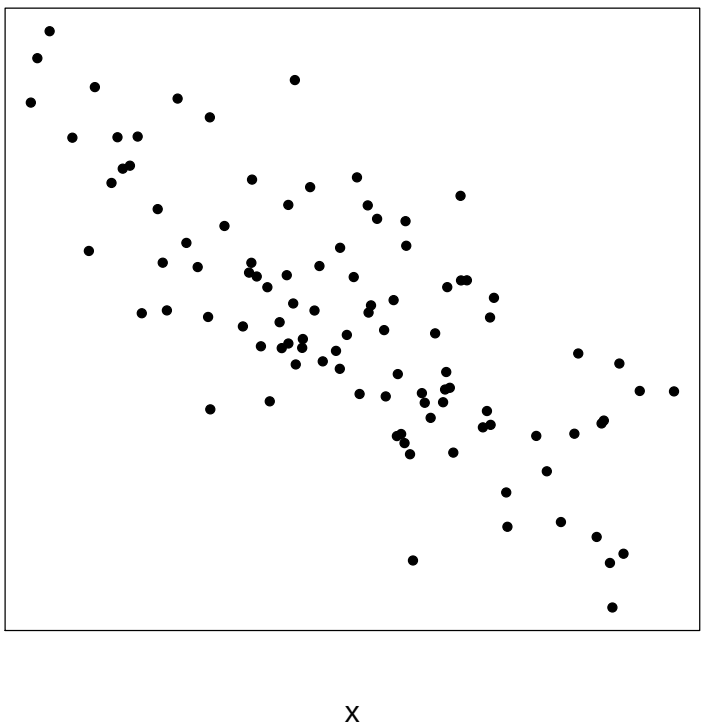



- Whole data set: $r = -0.35$




- Excluding China: $r = -0.02$

# Group Exercise




**What is your best guess of the correlation between $x$ and $y$?**

1. $-1$
2. $-0.7$
3. $-0.2$
4. $0$
5. $0.2$
6. $0.7$
7. $1$

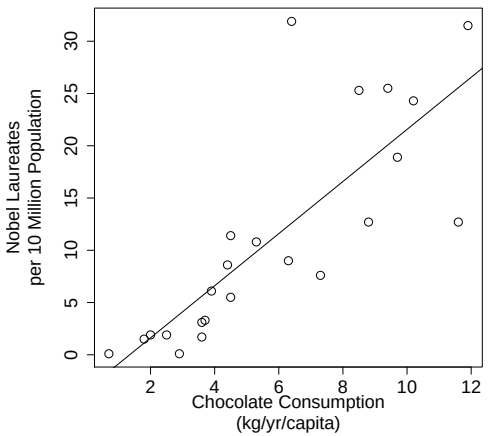
The Data

Correlation

Simple Linear Regression

Inference for SLR

# The idea




Estimate the line of best fit:
$\hat{y} = b_0 + b_1 x$

Correlation allows us to:

1. identify the direction of the association between $x$ and $y$
2. quantify the strength of the association between $x$ and $y$

A line of best fit additionally allows us to:

1. make predictions of $y$ based on $x$
2. further describe the relationship between $x$ and $y$ with the slope $b_1$

# Identifying the line

```
> lm(NEJM$nobel_rate~NEJM$chocolate)

Call:
lm(formula = NEJM$nobel_rate ~ NEJM$chocolate)

Coefficients:
   (Intercept)  NEJM$chocolate
        -3.400           2.496
```

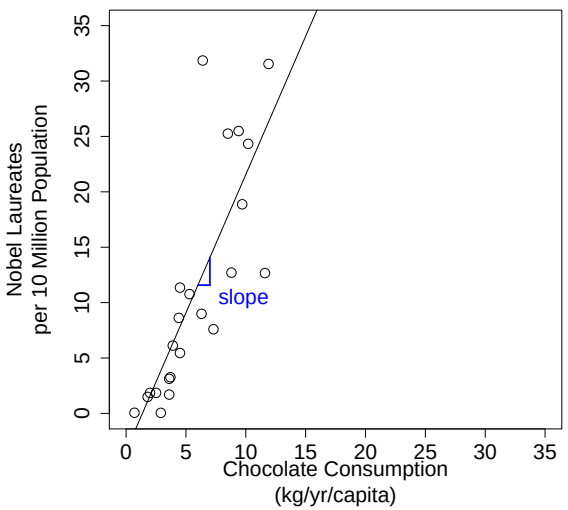
General form: $\hat{y} = b_0 + b_1 x$

Here, this is: $\hat{y} = b_0 + b_1 \times chocolate$

Estimated line: $\hat{y} = -3.4 + 2.5x$

Here, this is: $\hat{y} = -3.4 + 2.5 \times chocolate$

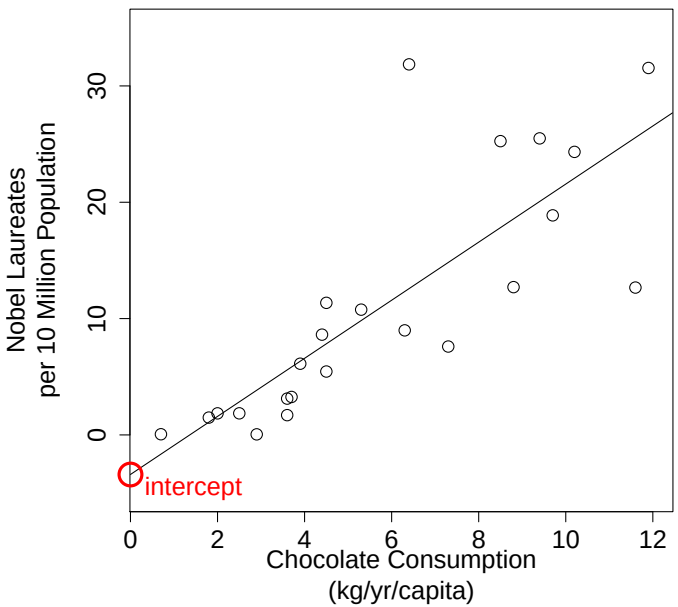
# The slope ($b_1$)




$$\hat{y} = -3.4 + 2.5x$$

- the slope of this line is 2.5
- the slope is the amount that $\hat{y}$ changes when $x$ increases by one unit
- for each additional 1kg/yr/capita of chocolate that a country consumes, the rate of nobel laureates per 10 million persons increases by 2.5
- the sign of the slope (positive or negative) indicates the direction of the association

# The $y$-intercept ($b_0$)




$$\hat{y} = -3.4 + 2.5x$$

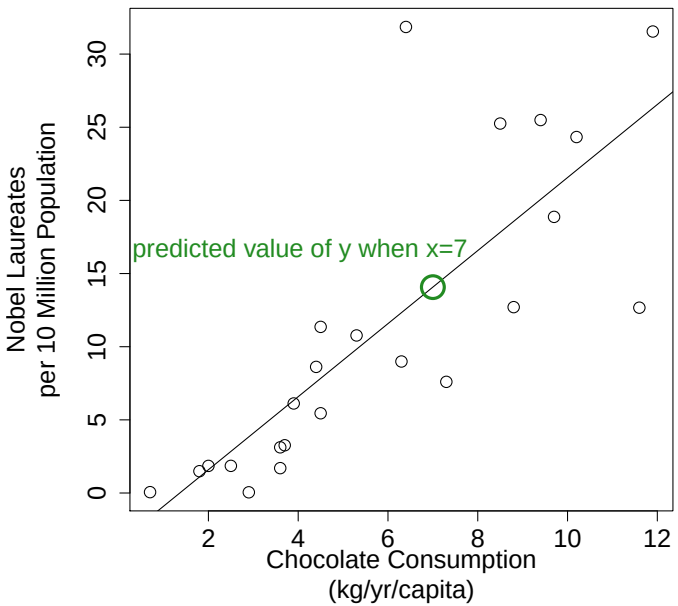
- the intercept of this line is -3.4
- the intercept is the predicted value of $y$ when $x = 0$
- when a country consumes no chocolate, its predicted number of Nobel laureates is -3.4
- intercepts aren't always interpretable

Suppose we use literacy rate of a country (measured as a percent) to predict the life expectancy of that country (measured in years), and we estimate the regression line to be $\hat{y} = 11.8 + 0.70x$.

**What is the interpretation of the slope?**

1. For each one percent increase in literacy rate, life expectancy increases by 0.7 years.
2. For each one year increase in life expectancy, literacy rate increases by 0.7 percentage points.
3. For each one percent increase in literacy rate, life expectancy increases by 11.8 years.
4. For each one year increase in life expectancy, literacy rate increases by 11.8 percentage points.
5. The predicted life expectancy of a country with a literacy rate of 0 is 11.8 years.
6. The predicted literacy rate of a country with a life expectancy of 0 is 11.8%.

# Prediction ($\hat{y}$)




$\hat{y} = -3.4 + 2.5x$

- when $x = 7$,
  $\hat{y} = -3.4 + 2.5 \times 7 = 14.1$
- the predicted number of Nobel Laureates for a country that consumes 7 kg/yr/capita of chocolate is 14.1

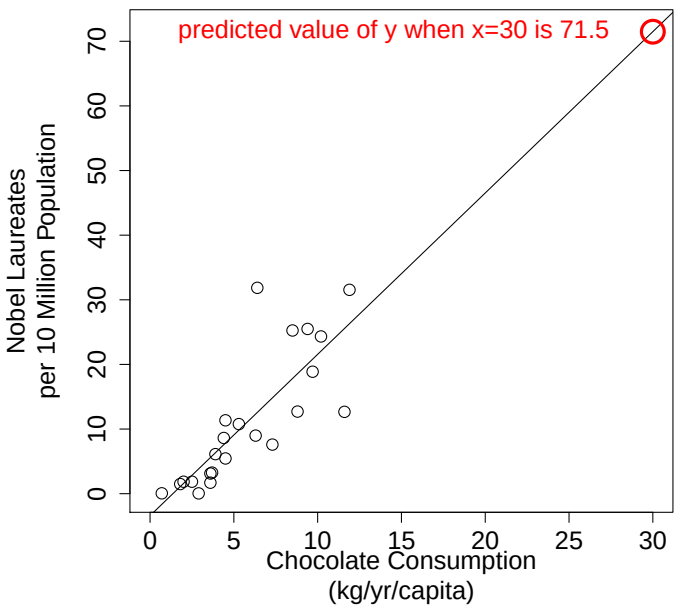
Using length of pregnancy in days (the gestation period) to predict babies' birth weight in pounds in the babies data set:

```
> cor(babies$bwt_lbs,babies$gestation)
0.42
> lm(bwt_lbs~gestation, data=babies)
Call:
lm(formula = bwt_lbs ~ gestation, data=babies)
Coefficients:
   (Intercept)        gestation
      -1.30920          0.03143
```

**What is the correct formula to predict the birth weight of a baby with a gestation period of 300 days?**

1. $\hat{y} = 0.03 - 1.31 \times 300$
2. $\hat{y} = 0.03 + 0.42 \times 300$
3. $\hat{y} = -1.31 + 0.03 \times 300$
4. $\hat{y} = -1.31 + 0.42 \times 300$

# Don't extrapolate!




- Extrapolation is using a regression line to predict $y$-values for $x$-values outside the observed range of the data.
- Extrapolation gets riskier the farther we move from the range of the given $x$-values.
- There is no guarantee that the relationship given by the regression equation holds outside the range of sampled $x$-values.

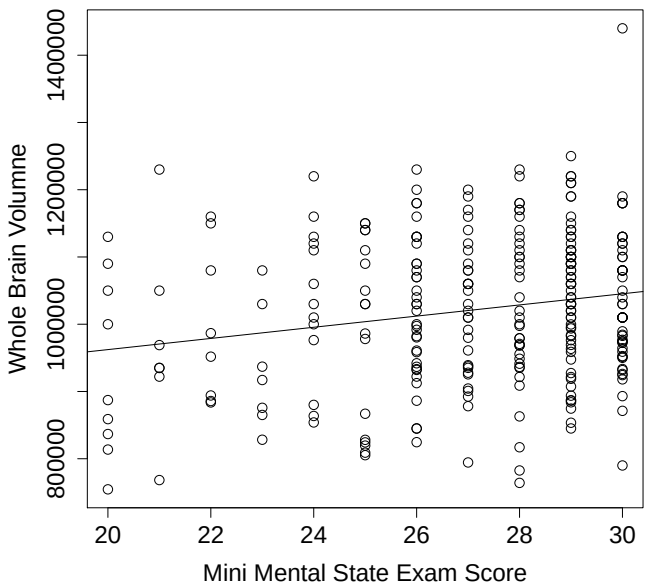
# Cautions in Regression

- Don't extrapolate
- The analysis is not robust to outliers (can affect estimates of correlation, slope, and intercept)
- Correlation/association does not imply causation
- Other variables can influence the analysis through confounding and interaction - you can control for this in *multivariable regression* where you use more than one variable in your model to predict $y$

**When using a simple linear regression model, which of the following allows us to assess the *association* between $x$ and $y$?**

1. the intercept ($b_0$)
2. the slope ($b_1$)
3. both
4. neither

Using MMSE score to predict whole brain volume in the ADNI data set: $r = 0.19$, $\hat{y} = 796239 + 8302 \times MMSE$




**What can we conclude about the strength of the association between MMSE and whole brain volume?**

1. there is a strong association because the slope is large
2. there is a strong association because the correlation is large
3. there is a weak association because the slope is small
4. there is a weak association because the correlation is small

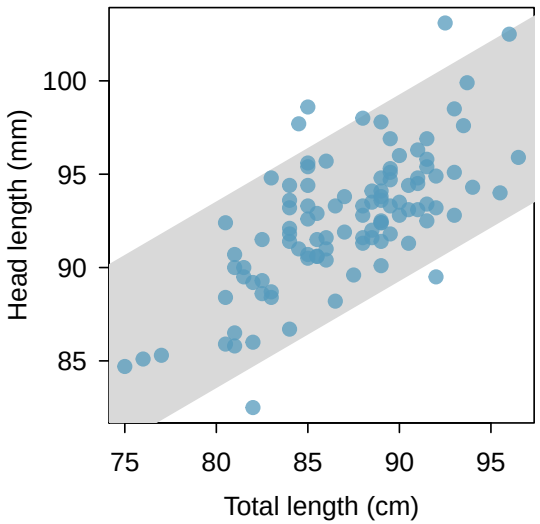
# The idea

- the *slope* of the line allows us to assess whether or not there is an association between $x$ and $y$
- a slope of 0 indicates no association between $x$ and $y$; a non-zero slope indicates some sort of association between $x$ and $y$
- we are estimating a slope $b_1$ based on sample data, but we want to draw conclusions about the slope in the population ($\beta_1$)
- We can do this with:
  1. a hypothesis test of $H_0$: $\beta_1 = 0$ vs $H_a$: $\beta_1 \neq 0$
  2. a confidence interval for $\beta_1$

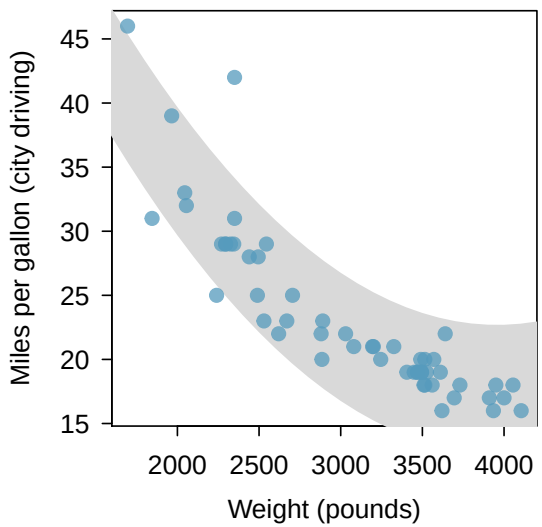
# Conditions for inference in SLR

1. observations are independent
2. linear relationship between $x$ and $y$
3. "normality" - not addressed in STAT 217 (requires residuals)
4. constant variability in $y$ about the regression line

# Example Figures




- possum data
- the spread in $y$ (head length) is about the same for all values of $x$ (total length)
- this satisfies "constant variability in $y$ about the regression line"




- car data
- the relationship between $x$ (weight) and $y$ (miles per gallon) follows a curve rather than a straight line
- this violates "linear relationship between $x$ and $y$"

Figure A Figure B Figure C Figure D

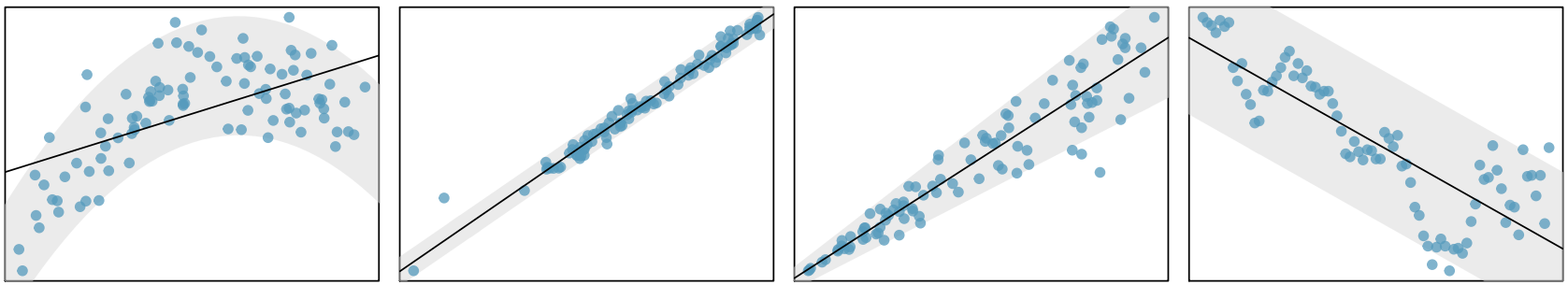
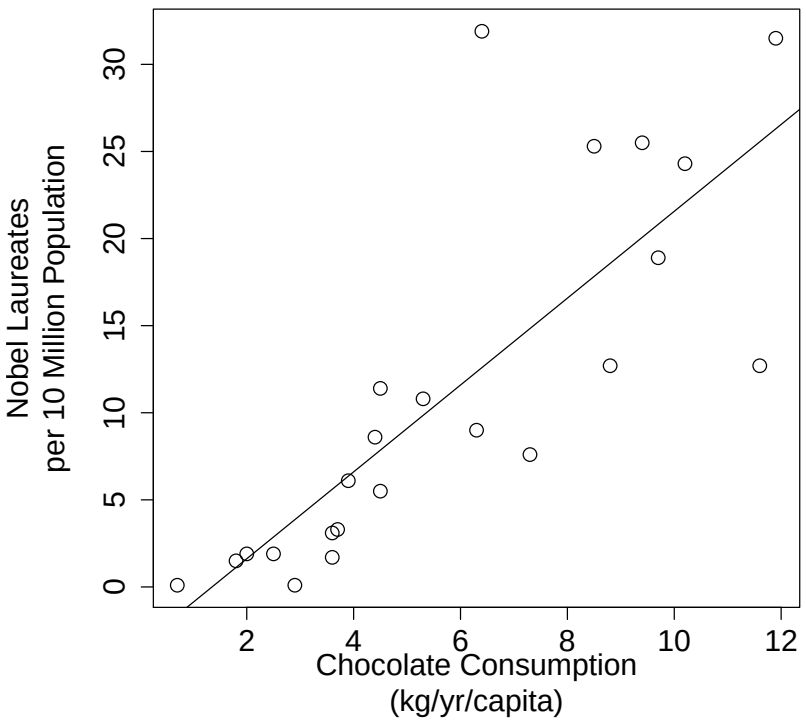


Figure (A/B/C/D) violates the independence condition.
Figure (A/B/C/D) violates the linear condition.
Figure (A/B/C/D) violates the constant variability condition.

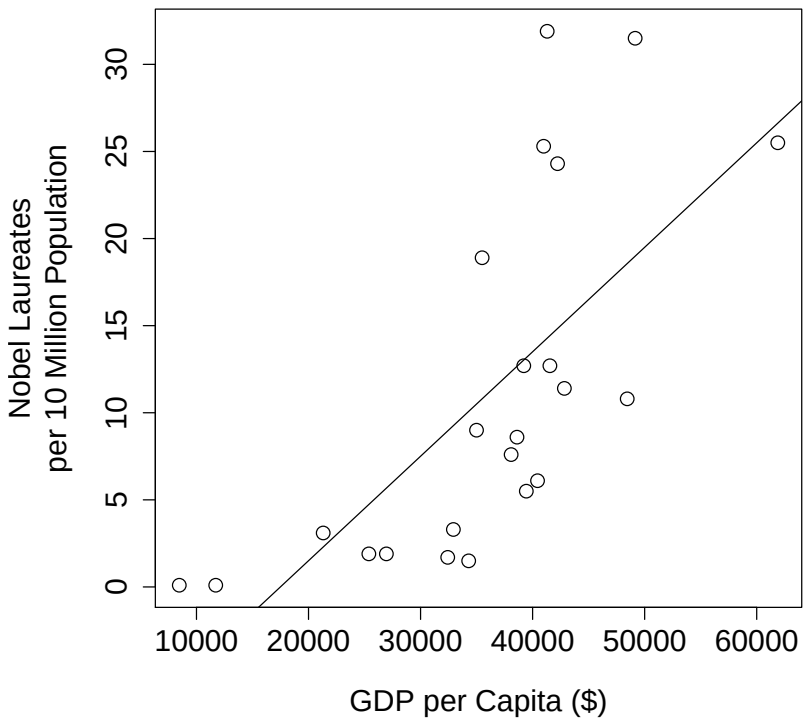
# Assessing conditions




**Which condition may be violated? Select the *best* answer.**

1. independence of observations
2. linear relationship between $x$ and $y$
3. constant variability in $y$ about the regression line
4. no conditions are violated

# Assessing conditions, continued




**Which condition may be violated? Select the *best* answer.**

1. independence of observations
2. linear relationship between $x$ and $y$
3. constant variability in $y$ about the regression line
4. no conditions are violated

# R results

```
> m1<-lm(NEJM$nobel_rate~NEJM$chocolate)
> summary(m1)

Call:
lm(formula = NEJM$nobel_rate ~ NEJM$chocolate)

Residuals:
    Min      1Q  Median      3Q     Max
-12.888  -2.953  -0.213   1.992  19.279

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3.400      2.699  -1.260    0.222
NEJM$chocolate    2.496      0.407   6.133 4.37e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.26 on 21 degrees of freedom
Multiple R-squared:  0.6418,        Adjusted R-squared:  0.6247
F-statistic: 37.62 on 1 and 21 DF,  p-value: 4.374e-06
```

# Focus on the coefficients box

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3.400      2.699  -1.260    0.222
NEJM$chocolate    2.496      0.407   6.133 4.37e-06 ***
```

On the `NEJM$chocolate` line:

**Hypotheses:** $H_0$: $\beta_1 = 0$ vs $H_a$: $\beta_1 \neq 0$

**Test statistic:** $t = \dfrac{\hat{\beta}_1 - 0}{se_{\hat{\beta}_1}} = \dfrac{2.496 - 0}{0.407} = 6.133$

***p*-value:** two-tailed area from $t$ distribution with $df = n - 2$; $p$-value$= 0.00000437$

**Conclusion:** At $\alpha = 0.05$ reject $H_0$; we have evidence of a positive association between chocolate consumption and rate of nobel laureates.

# Confidence interval for the slope

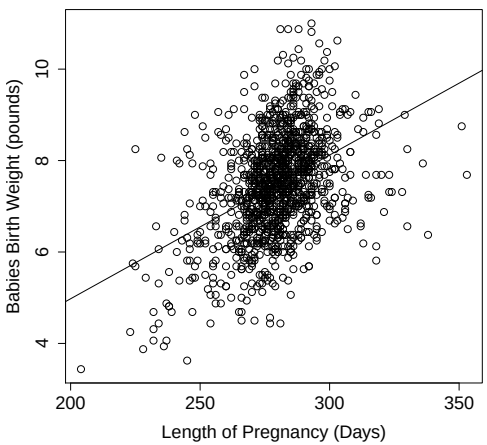
```
> confint(m1)
                    2.5 %   97.5 %
(Intercept)     -9.013147 2.212415
NEJM$chocolate   1.649869 3.342646
```

$$\hat{\beta}_1 \pm t^*_{df=n-2} \times se_{\hat{\beta}_1}$$
$$2.496 \pm 2.080 \times 0.407$$

- The 95% CI for $\beta_1$ is (1.65, 3.34)
- Because this CI is entirely positive, there is a *positive* association between rate of nobel laureates and chocolate consumption, such that as chocolate consumption increases so does the rate of nobel laureates.
- At the 95% confidence level, for each 1kg/yr/capita increase in chocolate consumption the rate of nobel laureates per 10 million persons increases by as few as 1.65 or as many as 3.34

```
> cor(babies$bwt_lbs,babies$gestation)
0.42

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.309203   0.540712  -2.421   0.1560
babies$gestation  0.031433   0.001931  16.274   <2e-16 ***
```

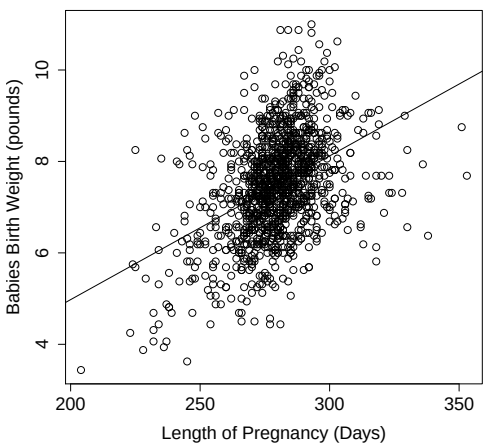



**Identify a number that represents _____ between birth weight and gestation days:**

1. the magnitude of the association
2. the strength of the evidence
3. the strength of the association

```
> cor(babies$bwt_lbs,babies$gestation)
0.42

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.309203   0.540712  -2.421   0.1560
babies$gestation  0.031433   0.001931  16.274   <2e-16 ***
```

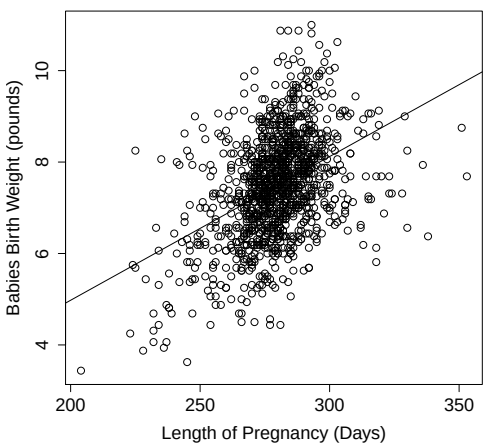



At $\alpha = 0.05$, what can we conclude? We (do/do not) have evidence that $\beta_1$ differs from zero; so we (do/do not) have evidence of an association between gestation days and birth weight.

1. do; do
2. do not; do not
3. do; do not
4. do not; do

```
> cor(babies$bwt_lbs,babies$gestation)
0.42

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.309203   0.540712  -2.421   0.1560
babies$gestation  0.031433   0.001931  16.274   <2e-16 ***
```




**Is there a strong linear association between gestation days and birth weight?**

1. Yes, because the $p$-value for $H_0$: $\beta_1 = 0$ is less than 0.05.
2. No. Although the $p$-value for $H_0$: $\beta_1 = 0$ is less than 0.05, the estimated slope is small.
3. No, because the $p$-value for $H_0$: $\beta_1 = 0$ is greater than 0.05.
4. Yes, the correlation is positive.
5. No, the correlation isn't that big.